第 40 卷 第 4 期 厦门大学学报(自然科学版) Vol. 40 No. 4
2001 年 7 月 Journal of Xiamen University (Natural Science) Jul. 2001

• 研究简报 •

文章编号: 0438-0479(2001)04-1000-04

一种基于粗糙集的单一属性分类的约简方法

康胜武, 曾 一锋, 王华火, 王应明
(厦门大学自动化系, 福建 厦门 361005)

摘要: 粗糙集的应用中, 对象集通常具有数量大、属性多、单一属性分类大的特点, 这是在已有知识基础上提出规则时所遇到的 3 个主要问题. 针对其中单一属性分类的约简问题提出了一种新的方法, 它采用了先合并后分解再综合的思想简化问题的求解, 能快速有效地发掘信息中蕴涵的规则.

关键词: 粗糙集; 属性约简; 属性核值表
中图分类号: TP 18 文献标识码: A

1 粗糙集

粗糙集理论是由波兰科学家提出的一种处理含糊和不精确性问题的新型数学工具. 它能有效地分析不精确, 不一致, 不完整等各种不完备的信息, 还可以对数据进行分析和推理, 从中发现隐含的知识, 揭示潜在的规律. 它的主要思想就是根据目前已有的对给定问题的知识, 将问题的论域进行划分, 在保持分类能力不变的前提下, 通过知识约简, 导出概念的分类规则.

1.1 基本概念

知识表达系统可表示为 $S = < U, C, D, V, f >$, U 是对象的集合, $A = C \cup D$ 是属性集合, C 和 D 分别为条件属性集和决策属性集, $V = \bigcup_{a \in A} a$ 是属性值的集合, V_a 表示属性 $a \in A$ 的属性值范围, $f: U \times A \rightarrow V$ 是一个信息函数, 它指定 U 中每一个对象 x 的属性值. 定义 $IND(C)$ 和 $IND(D)$ 是 U 上关于属性 C 和 D 的两个不可辨认关系, 并满足: $x, y \in U$ 且 x, y 关于 C 等价, 即对所有的 $a \in C$, $(x, y) \in IND(C)$ 当且仅当 $f(x, a) = f(y, a)$.

1.2 属性约简和值约简

在一个决策系统中, 决策属性和条件属性往往存在某些程度上的依赖和关联. 属性约简就是在保持分类能力不变的前提下, 求出条件属性和决策属性之间的最小依赖关系. 常用的有文献[1] 中的方法和可辨识矩阵方法[2].

得到最佳属性约简后, 可对决策表进行最小值约简, 即找出每条记录中对决策影响最大的

收稿日期: 2000-05-09
基金项目: 高等学校青年骨干教师基金(教计师(2000)65) 和福建省自然科学基金(A0010002) 资助项目
作者简介: 康胜武(1973–), 男, 硕士研究生.
© 1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://w

属性值, 从而得到属性核值表, 最后经过整理就得到所需的规则. 具体方法见文献[3, 4] .

1.3　单一属性约简

在具体问题中, 通常希望属性的等价类不要过多, 这样易于问题的处理, 所以我们可以先对某些属性进行适当的约简. 文献[5] 提出了一种多层次, 逐步求精的发掘算法. 每个属性对应一个概念层次树, 离树根越近的节点层次越高, 形成的等价类数量(单一属性分类) 就越少, 表示概念的泛化程度也较高. 树分得越细, 等价类数量就越多, 最后得出的规则也越精细. 可以看出, 概念层次树实质上是在较高层次上对单一属性分类进行了约简, 但所研究的属性本身具有明显的层次性, 如所举例中的学历属性, 但对于不存在层次性的属性分类, 这时就很难建立层次树.

我们可以在根节点层, 对节点进行归并, 然后把原表分解为相容与不相容两个决策表, 分别求出两个规则集, 最后对两个规则集进行综合, 就得到最终所需的结果, 即先把复杂问题分解为两个较容易问题的求解, 然后再综合的思想.

具体算法如下:

设A 为条件属性集合, D 为决策属性集合, $f(x_{ij})$ 为条件属性 a_i 的等价类 X_j 所对应的决策值. $U/a_i, \{X_{ij}\}$, i 为条件属性个数, j 为条件属性的等价类数.

第 1 步: 确定要合并的等价类 X_{ij}, 遵循以下原则:

1) 对 $a_i \in A$, $f(X_{ij})$ 为单一值, 可考虑把 X_{ij} 并入其他等价类 X_{ij} 中.

2) 若 $\dfrac{card\left(X_{ij_1}\ f(X_{ij_1}) = \underset{d\in D}{V_d}\right)}{card(X_{ij_1})} = \dfrac{card\left(X_{ij_2}\ f(X_{ij_2}) = \underset{d\in D}{V_d}\right)}{card(X_{ij_2})}$, 等式两边 V_d 取值相同, 可把 X_{ij_1} 与 X_{ij_2} 归为一类.

第 2 步: 检查等价类合并后的新表, 合并相同的记录, 并且把新表分解为相容与不相容两个表.

第 3 步: 求出相容表的规则集 S1, 具体方法见文献[6] , 把不相容表中的变动的等价类重新还原为原来的类, 求出规则集 S2.

第 4 步: 对 S1 和 S2 进行综合, 步骤如下:

1) 在 S1 中, 找出这样的规则, 其条件属性值在不相容表中无相应匹配值.

2) 在 S2 中, 找出可由变更条件属性唯一确定的规则.

3) 在 S1 和 S2 剩余规则中, 根据规则确定性原则. 按文献[1] 中的方法, 先还原标记为? 的属性, 找出属性值与决策值存在一一映射的规则, 然后同理处理标记为×的属性.

第 5 步: 综合第 4 步中得到的规则, 得所需的结果.

2　举　例

现以一气象状况实例作为对象集, 如表 1 所示. 其中: a1, a2, a3, a4 为条件属性, 数字分别代表:

Outlook(a1) : 1—sunny; 2—overcast; 3—rain

Temperature(a2) : 1—hot; 2—mild; 3—cool

Humidity(a3) : 1—high; 2—normal

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://w

表 1 气象状况表

Tab. 1 The status of weather

U	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Outlook(a1)	1	1	2	3	3	3	2	1	1	3	1	2	2	3
Temperature(a2)	1	1	1	2	3	3	3	2	3	2	2	2	1	2
Humidity(a3)	1	1	1	1	2	2	2	1	2	2	2	1	2	1
Windy(a4)	1	2	1	1	1	2	2	1	1	1	2	2	1	2
Class(d)	N	N	P	P	P	N	P	N	P	P	P	P	P	N

Windy(a4): 1—false; 2—true

d 为条件属性

用可辨识矩阵方法得到的规则如下:

1) If (a1, sunny) and (a3, high) then class= N

2) If (a1, overcast) then class= P

3) If (a1, rain) and(a4, false) then class= P

4) If (a1, rian) and(a4, true) then class= N

5) If (a1, sunny) and (a3, normal) then class= P

现在我们用本文的方法来求解.

第 1 步: 把 a1 属性中的 2 归入到 3 中, 把 a2 属性中的 3 归入到 2 中, 并且把变化得到的新表分解为相容和不相容 2 个决策表, 如表 2, 3 所示;

表 2 相容决策表

Tab. 2 The consistent decision table

U	a1	a2	a3	a4	d
1	1	1	1	1	N
2	1	1	1	2	N
3	3	1	1	1	P
4	3	2	1	1	P
5, 10	3	2	2	1	P
8	1	2	1	1	N
9	1	2	2	1	P
11	1	2	2	2	P
13	3	1	2	1	P

第 2 步: 由相容决策表得到的属性约简及核值表如表 4 所示.

表 3 不相容决策表

Tab. 3 The inconsistent decision table

U	a1	a2	a3	a4	d
6	3	2	2	2	N
7	3	2	2	2	P
12	3	2	1	2	P
14	3	2	1	2	n

表 4 核值表

Tab. 4 The core value table

a1	a3	D		
1	1	N	1	1
3	1	P	3	*
3	2	P	*	*
1	2	p	*	2

可得到的确定性规则有 1 条, 即:

1) If (a1, sunny) and (a3, high) then class= N.

把不相容决策表中的变动的等价类重新还原为原来的类而得到的表 5.

表 5 不相容决策表的还原表

Tab. 5 The restorable table from the inconsistent table

U	a1	a2	a3	a4	d
6	3	3	2	2	N
7	2	3	2	2	P
12	2	2	1	2	P
14	3	2	1	2	N

由表 5 得到的属性约简及核值表如表 6.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://w

由表得到的确定性规则有:

2) If (a1, Overcast) then class= P

第 3 步: 在表 4 和表 6 剩余规则中根据确定性原则找出其它规则:

3) If(a1, rain) and (a4, false) then class= P

4) If(a1, rian) and (a4, true) then class= N

5) If((a1, sunny) and (a3, normal) then class= P

可看出所得结果与可辨识矩阵方法是一致的.

表 6　表 5 的核值表

Tab. 6　The core value table of the Tab. 5

a1	a4	d		
3	2	N	3	*
2	2	P	2	*

3　结束语

目前大多数是讨论对条件属性的约简, 即所谓的属性约简, 对于单一条件属性约简问题的研究还很少. 本文提出的方法对于单一条件属性分类较多的情况下, 能有效使问题的求解得到简化. 但本文仅讨论了条件属性分类中的一些特殊情况, 并且决策属性为确定和不确定型, 对于条件属性中的其它形式和多决策属性的情况, 能否提出更一般的模型还有待进一步的研究.

参考文献:

[1] 吴福保, 李奇, 宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法[J]. 控制与决策, 1999. 03: 206 – 211.

[2] Skowron A, suraj Z. Discovery of concurrent data models from experimental data tables: A Rough set approach[R]. Institute of Computer Science, Warsaw University of Technology, Research Report: 1995.

[3] 常犁云, 王国胤, 吴渝. 一种基于 RoughSet 理论的属性约简及规则提取方法[J]. 软件学报, 1999, 10 (11): 1 206– 1 211.

[4] 曾黄麟编著. 粗集理论及其应用: 关于数据推理的新方法[M]. 重庆: 重庆大学出版社, 1996.

[5] 刘发升, 杨炳儒. 一种基于粗糙集的多层次逐步求精的发掘算法[J]. 计算机工程与应用, 1999. 5: 11–12.

[6] Pawlak Z. RoughSet, Theoretical Aspects of Reasoning About Data[M]. Warsaw: Klumer Academic Publisher, 1992.

[7] Pawlak Z. Rough set[J]. Intern. J. of Comp. and Inform. Sci., 1982, 11(5): 341– 356.

An Reduction Approach for Single Attribute Class Based on Rough Sets Theory

KANG Sheng-wu, ZENG Yi-feng, WANG Hua-huo, WANG Ying-ming

(Dept. of Automation, Xiamen Univ., Xiamen 361005, China)

Abstract: During the application of the Rough sets, the object sets usually have a character of large quantity, attributes and single attribute class. These problems appear when the regulation is proposed based on the past knowledge. With the view of single attribute class, this paper proposes a new method which is combining class firstly, then discomposing, last synthesizing. This way can simplify the solution and mine the regulation in the information quickly and sufficiently.

Key words: rough sets theory; attribute reduction; value reduction

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://w